

**КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ ФУНДАМЕНТАЛЬНОЙ МЕДИЦИНЫ И
БИОЛОГИ**

Кафедра биохимии и биотехнологии

Н.И.АКБЕРОВА

**АНАЛИЗ ДАННЫХ СЕКВЕНИРОВАНИЯ
ТРАНСКРИПТОМА И МЕТАБОЛОМА**

Учебно-методическое пособие

Казань – 2014

Секвенирование : RNA-SEQ и метагеномика

[необходимый софт: доступ к Интернету]

RNA-Seq или секвенирование транскриптома (Whole Transcriptome Sequencing “WTS”)

RNA-Seq - изучение транскрибируемых генов организма с помощью секвенирования следующего поколения РНК из интересующих образцов. Обычно РНК обратнo транскрибируется в кДНК перед секвенированием. РНК-Seq является количественной оценкой профиля транскрипции: можно определить, какие гены «включены», а также относительные уровни транскрипции. Это позволяет выявить функциональную активность в данном образце: определять аллель-специфическую экспрессию генов и находить новые транскрипты.

Сравнение RNA -Seq образцов, взятых из различных тканей или взятых из того же источника в различных условиях, позволяет исследовать дифференциальную экспрессию генов в ответ на программу развития или изменения окружающей среды (выявлять, какие гены "включены" или "выключены" в этих условиях, как гены регулируются в данных условиях.

В 1. RNA-Seq анализ

RNA-Seq анализ можно рассматривать как расширение много лет используемых методов, таких как ESTs, SAGE, and MPSS (expressed sequence tags, serial analysis of gene expression, and massively-parallel signature sequencing, соответственно) для анализа экспрессии генов. Главное отличие в том, что общее число "тегов", которые создаются для транскрипта данной популяции намного выше, в связи с эффективностью NGS-машин, которые позволяют провести секвенирование дешевле, но с повышенной точностью и чувствительностью.

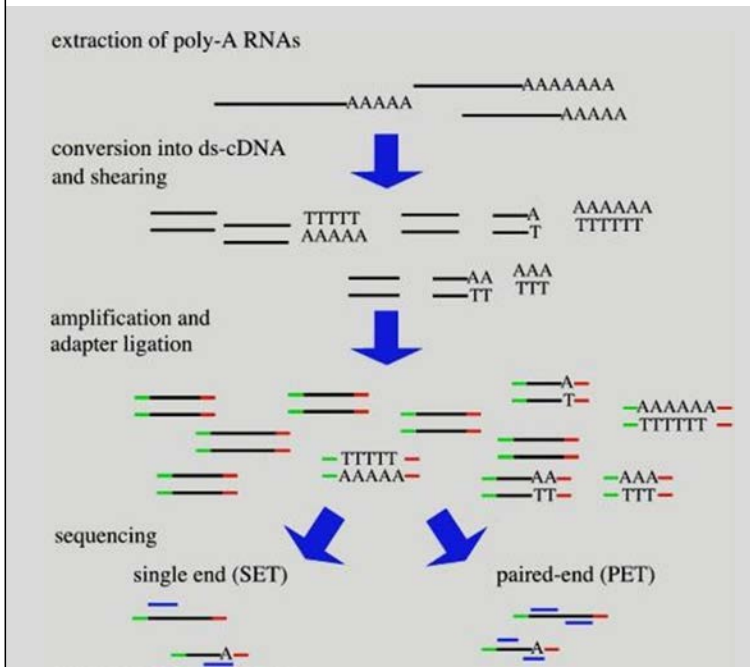


Рис. Протокол RNA -Seq анализа. РНК экстрагируют из интересующей пробы, обратно транскрибируют в кДНК, нарезают и преобразуют в библиотеки секвенирования для Roche 454 или Illumina и секвенируются. Секвенирование парных концов (PET) включает в себя несколько более сложный шаг подготовки библиотек, но позволяет лучше осуществлять сборку конечных последовательностей. (Взято с <http://cmb.molgen.mpg.de>)

С учетом всех этих способов, мРНК преобразуется в кДНК с помощью шага обратной транскрипции. В случае RNA-Seq, двухцепочечную молекулы кДНК нарезают на фрагменты меньшего размера, к которым лигируют адаптеры. Фрагменты затем секвенируют, либо с помощью single end reads или paired-end методологии, в которой генерируются короткие ряды, меченые с обоих концов. Поскольку расстояние между концами примерно известно, в последнем случае может быть достигнуто лучшее, неоднозначное картирование ридов

В любых способах профилирования экспрессии «нормализация» - важная процедура, чтобы можно было сравнивать оценки уровня экспрессии из разных экспериментов (т.е., мы говорим, что гены дифференциально экспрессируются, тогда как на самом деле у нас физически разное количество входного материала). Для RNA-Seq типична нормализация "Fragments Per Kilobase of exon per Million fragments mapped (FPKM)", или первоначально "Reads Per Kilobase of exon per Million fragments mapped (RPKM)" Заметим, что нормализация в таком контексте не относится к традиционному статистическому определению. Обратите внимание, что "нормализация" в этом контексте не относится к традиционным статистическим определением масштабирования все численных переменных в диапазоне [0,1].

Мы рассмотрим два различных набора данных RNA -Seq. Первый набор данных получали из вида риса *Oryza glaberrima*. Это африканской культивируемый вид риса, геном которого был секвенирован, что позволяет все RNA -Seq риды, полученные с использованием образцов РНК из него, откартировать на геном и идентифицировать.

Второй набор данных из *Arabidopsis thaliana*.

Эксперимент с рисом включал проведения RNA -Seq на различных тканях в растения риса, что позволило определять, как уровни экспрессии генов меняются в зависимости от физического расположения ткани в растении (лист, корень и т.д.).

1. Войдите на страницу Plant MPSS databases <http://mpss.udel.edu>.
2. Перейдите вниз к базе данных по рису (Rice databases), нажмите на [link](#) в столбце RNA-Seq для организма “Rice_glab”, и попадете на страницу http://mpss.udel.edu/rice_glab_RNAseq
 - a) Сколько генов идентифицировано в геноме *Oryza glaberrima*?
 - b) Сколько хромосом имеет этот вид?
3. Нажмите на “Library Information” вверху страницы.
 - a) Сколько RNA-Seq библиотек доступны для этого вида?
 - b) Из каких тканей получены библиотеки? Для чего нужно было секвенировать разные ткани одного и того же растения с использованием RNA-Seq?
4. Запомните, какие имена у библиотек, и из каких тканей растения они были получены. Вернуться на главную страницу Rice Oglab (“Home/basic queries” link).

5. Кликните на одну из хромосом (например, Chromosome 4) , чтобы расширить возможность просмотра

- a) Что означают красные и синие полосы? Почему они окрашены по-разному? (подсказка: нажмите на “legend” ссылку вверху, чтобы увидеть легенду)).
- b) Что отражают розовые области?



Рис. 1. Последовательность хромосомы 4 *Oryza glaberrima*

6. Щелкните в любом месте на увеличенном участке хромосомы для увеличения на один шаг дальше.

- a) Что представляют из себя серые полосы?

Если вы не видите серых полос переместите вид вправо или влево на 1 или 2 МВ.

7. Чтобы посмотреть конкретный ген и его паттерны экспрессии, вернуться к вкладке Home/basic queries и произвести поиск "Orgla03g0398400" в

“Protein Entry code”. Используйте значения по умолчанию.

а) Что означают серые и белые звезды на графике?

б) Какова предсказанная функция этого гена ?

8. Щелкните по кнопке “RNA-seq libraries” в “View library abundance”.
Измените опции Control Panel на «display the libraries separately» (линк в
тексте вверху страницы).

а) Что изменилось? Что показывает такое представление?

б) Какой образец имеет наивысшую экспрессию?

Дает ли это смысл предсказанной функции гена?

Контрольный вопрос 1

Какой образец экспрессируется слабее? OgR, OgP or OgL? (Имейте в виду, что RuBisCO участвует в фотосинтезе, так где вы не ожидаете выражение RuBisCO?)

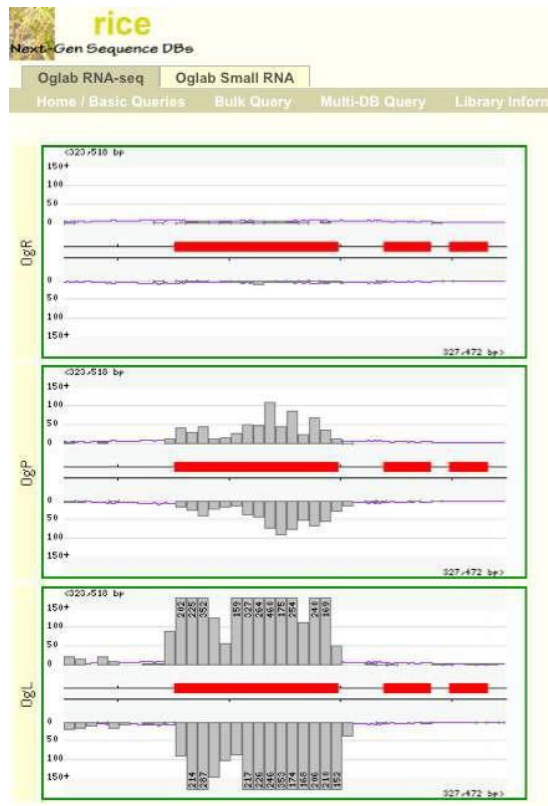


Рис. 2. Графическое представление RNA-Seq ридов из трех различных библиотек тканей, откартированное на интересующий ген.

9. Вернитесь к “tabular view” этого гена. Внизу страницы посмотрите “RNA-seq Information”

а) Что это за данные? Как они соотносятся с графиками, которые вы смотрели в пункте 8??

Второй набор данных RNA-Seq был получен Filichkin et al. (2009) на *Arabidopsis thaliana*, которая также имеет полную последовательность генома, что позволяет картировать RNA-Seq риды на геном. В этом случае, RNA-Seq информация была подготовлена для *Arabidopsis* в разных стрессовых условиях: засухи, засоления, повышенной освещенности, жары, холода. В эксперименте тестировались регуляция генов и сплайсинг генов в стрессовых условиях, чтобы увидеть, как экспрессия генов или реальная структура белков

изменялись в различных природных условиях. Эксперимент также позволил проверить известные структуры транскриптов. Сплайсинг это удаление интронов при созревании мРНК, которая затем служит в качестве шаблона для трансляции. Возможны несколько различных видов альтернативного сплайсинга: альтернативный сплайсинг в обоих интрон акцепторных и донорных сайтах сплайсинга, альтернативные переходы сплайсинга, и альтернативные интронные последовательности. Все такого рода события приводят к тому, что производится различные транскрипты, что в свою очередь может привести к производству различных белков.

Здесь мы рассмотрим пример альтернативного сплайсинга, где ген кодирует различные белки в зависимости от интронов и экзонов, включенных или исключенных во время транскрипции.

10. Откройте браузер и перейдите к <http://mockler-jbrowse.mocklerlab.org/jbrowse.athal/?loc=Chr2%3A12414112..12415692> (если переход по линку не работает, скопируйте и вставьте этот адрес в адресную строку браузера). Это геномный браузер для *Arabidopsis thaliana*, позволяющий увеличивать участки хромосом. Этот участок содержит ген, кодирующий белок аннотированный как “Outer Envelope Protein 16” (At2g28900).

11. В JBrowse дважды щелкните на следующем треки, чтобы добавить их в браузер: “Tair 10 Genome Annotation”, “Col0 Control RNA-Seq Coverage” и “Cold Stress RNA- Seq Coverage”. Вы также можете перетащить эти плитки в главную область браузера. Вы можете перемещать разделитель между панелями Available Tracks и основной части окна направо, чтобы иметь возможность увидеть полные имена треков. Обратите внимание, что треки не имеют определенного порядка.

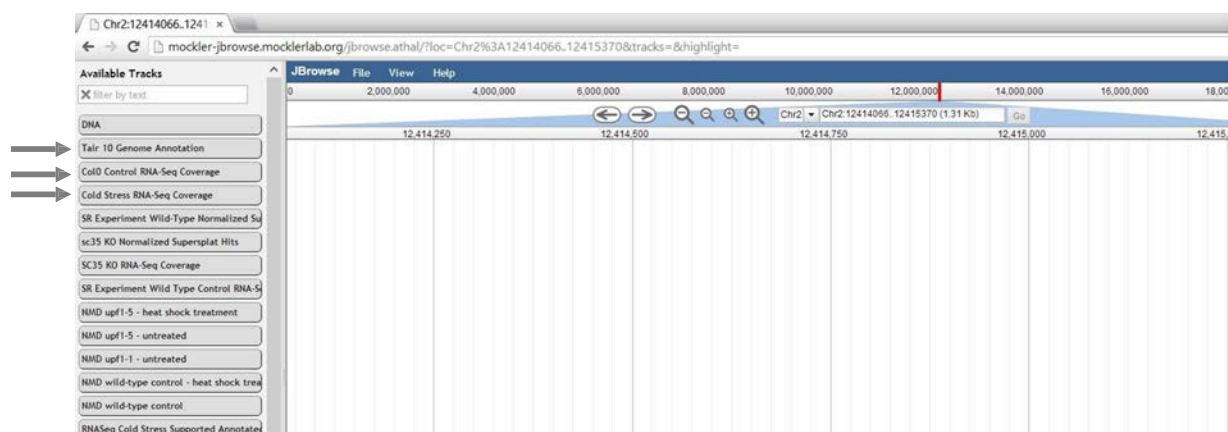


Рис. 3. Выбор трека для *Arabidopsis thaliana* RNA-Seq наборов данных. Выберите три набора данных, указанных стрелками: : “Tair 10 Genome Annotation”, “Col0 Control RNA-Seq coverage” и “Cold Stress RNA-Seq coverage”.

12. На горизонтальной оси графика показаны позиции нуклеотидов представляющего интерес гена. Для каждого выбранного RNA-Seq трека, RNA-Seq риды, картирующиеся на этом гене отображаются в виде гистограмм, показывающих плотность покрытия ридов данного конкретного нуклеотида. Чем выше гистограмма, тем больше ридов картируется здесь, и, следовательно, тем выше экспрессия этой части гена. Отметим, что данные, представленные здесь, были log-преобразованы и, чтобы получить фактическое количество ридов в данном положении на гистограмме, просто наведите курсор мыши на эту позицию. Пробелы в гистограмме показывают интроны или экзоны, что не были транскрибированы в мРНК анализируемого образца.

а) Сколько экзонов экспрессируются в контрольном наборе данных (зеленые на рис. 4)? Сравните это с моделью гена, которая показывает архитектуру гена с экзонами и UTR, которые

выделены красным цветом.

б) Как RNA-Seq профиль из Cold stress condition отличается от этого гена? На основе модели гена, как вы думаете, что происходит?

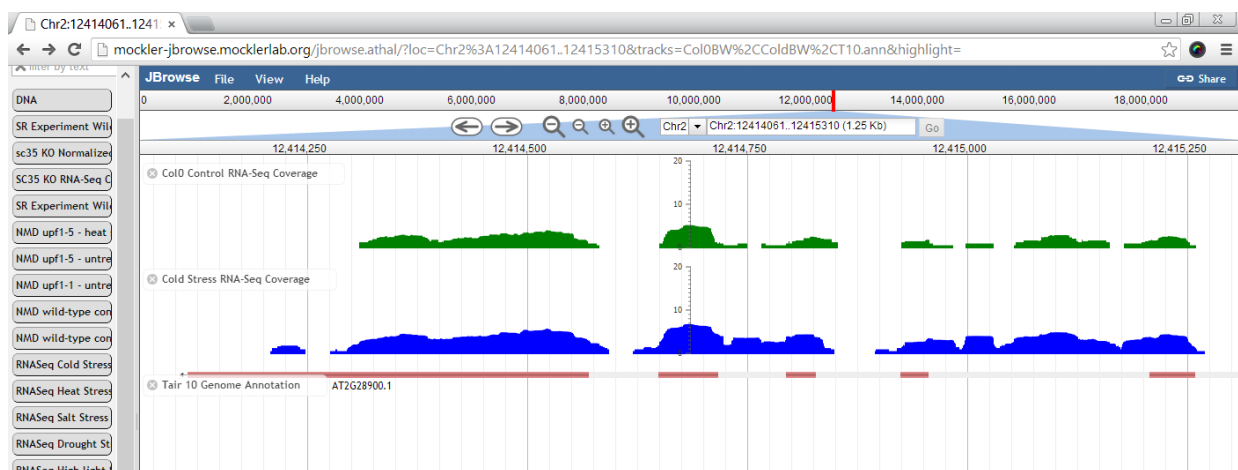


Рис. 4. JBrowse map of two RNA-Seq data sets on the Outer Envelope Protein 16, with CDS and gene model maps. The Col0 Control RNA-Seq has been set to green by editing the configuration file. JBrowse карта двух наборов данных RN

A -Seq белка внешней оболочки Outer Envelope Protein 16, с CDS и картой модельного гена. Col0 Control RNA-Seq был установлен на зеленый, путем редактирования конфигурационного файла

13. Добавьте еще несколько RNA -Seq треков, например, полученных для засухи или теплового стресса, перетаскив их на основную панель.

а) Видите ли вы доказательства различных альтернативных вариантов сплайсинга при измененных условиях? Какой вид альтернативного сплайсинга, вы думаете, имеет место?

Четкое определение событий альтернативного сплайсинга по всему геному вручную было бы проблемой, поэтому исследователи разработали алгоритмы, чтобы сделать это автоматически и статистически значимо, например, с помощью программы supersplat (Bryant et al., 2010, <http://www.ncbi.nlm.nih.gov/pubmed/20410051>) или TAU, описанной в статье Filichkin et al. (2009). Изучение RNA-Seq треков в геномных браузерах, тем не менее может быть информативными на основе ген - за- геном сравнения.

Метагеномика

Метагеномика изучает смешанные сообщества организмов путем секвенирования ДНК, извлеченной из сообщества в целом. Последовательность метагенома предоставляет информацию о том, какие виды содержит сообщество (кто в нем есть), а также о метаболических функциях этих видов (то, что они в состоянии сделать). Метагеномика предоставляет геномную, а не транскриптомную информацию; ген с предсказанной функцией, присутствующий в метагеноме, не обязательно экспрессироваться, но он присутствует в сообществе таким образом, может быть функционально важным.

Метагеномы полезны для сравнения изменений в составе сообществ с течением времени, для картирования РНК-Seq ридов или протеомных данных и для выявления новых генов.

В 2. Метагеномика

Термин метагеномика был введен в 1998 году Handelsman et al. (see [http://dx.doi.org/10.1016/S1074-5521\(98\)90108-9](http://dx.doi.org/10.1016/S1074-5521(98)90108-9)) и широко популяризирован новаторскими метагеномными исследованиями Крейга Вентера воды из якобы "мертвого" Саргассового моря вблизи Бермудских островов в 2004 г. (<http://dx.doi.org/10.1126/science.1093857>). Это исследование показало широкий спектр прокариотических - более 1800 – филотипов, присутствующих в этих водах. Метагеномика была определена как "применение методов современной геномики к изучению сообществ микроорганизмов непосредственно в их естественной среде, минуя необходимость выделения и лабораторного выращивания отдельных видов" (см <http://dx.doi.org/10.1371%2Fjournal.pcbi.0010024>). На рисунке 1 показана схема типичного метагеномного процесса.

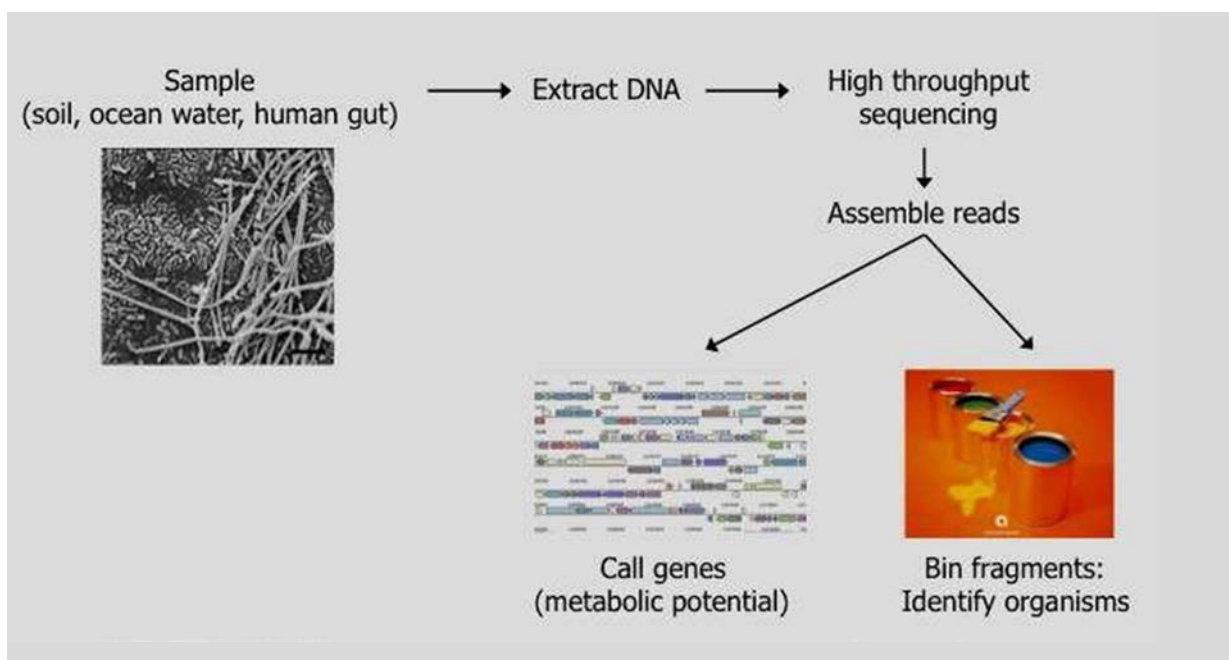


Рис 1: Протокол секвенирования метагенома. ДНК экстрагируют из смешанного сообщества для секвенирования спомощью NGS, возможно, со стадией фильтрации gJ размеру видов. Последовательность после сборки дает проект метаболического потенциала сообщества, а также таксономический профиль (кто есть, и то, что он потенциально способен делать). Изображение из JGI

В отличие от более ранних исследований, основанных на секвенировании только 16S рибосомных рРНК последовательностей, амплифицированных из проб окружающей среды, метагеномика также обеспечивает индикацию метаболического потенциала сообщества в плане метаболических промежуточных продуктов из одного вида, которые могут быть использованы другим видом, который сам не способен синтезировать этот продукт, а также причины, почему отдельные виды могут выжить в определенной нише (например, большое количество систем, поглощающих железо при низких рН среды, где железо не очень доступно). Число видов в сообществе влияет на количество целых геномов, которые могут быть собраны из коротких последовательностей ридов в проекте метагеномного секвенирования. Чем меньше число видов, тем больше последовательностей полных генома могут быть получены (рисунок 2).

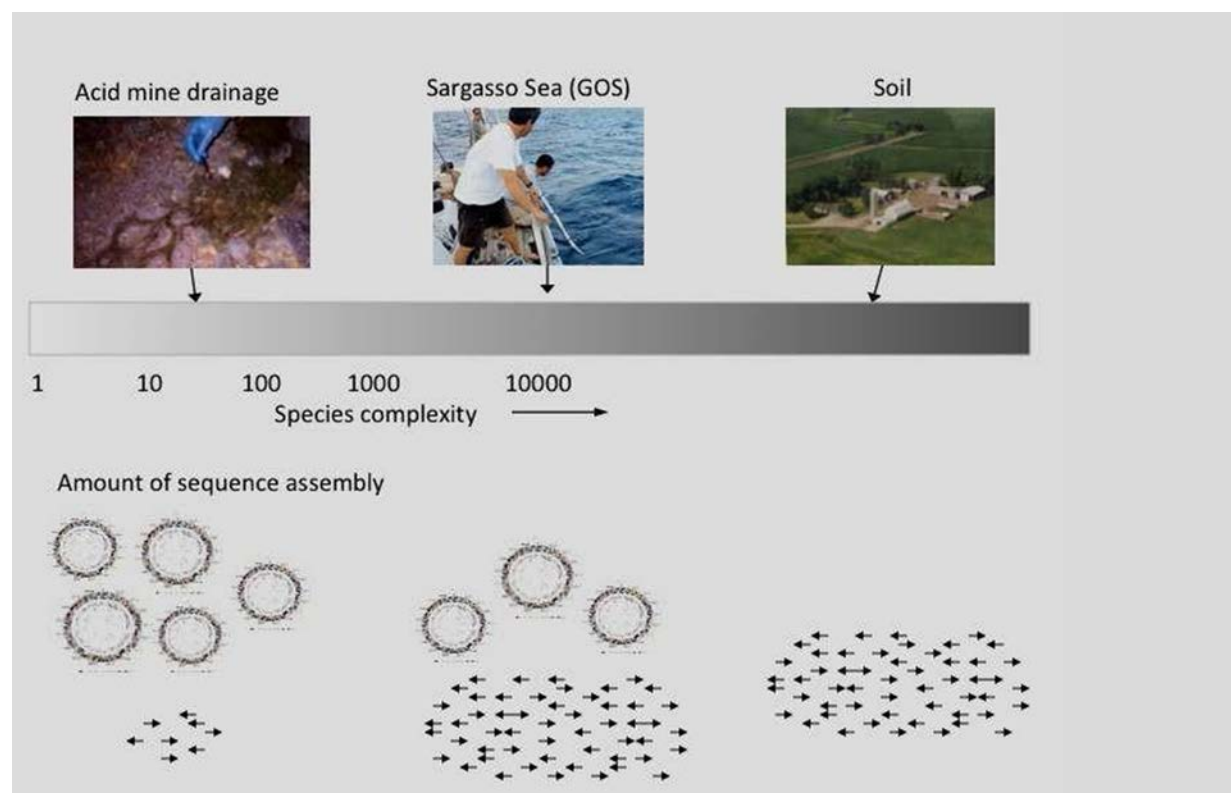


Рис 2: Как комплексность сообщества влияет на состав в метагенома. Менее сложные сообщества (например, биопленки кислотного шахтного дренажа) имеют лишь несколько доминантных организмов, чьи геномы собираются относительно полностью (круги в “Amount of sequence assembly”). Более сложные сообщества (например, почвы) собираются плохо, со многими фрагментами последовательностей, которые содержат только несколько генов. Изображение из JGI.

В этом упражнении мы рассмотрим два метагеномные последовательности с помощью веб-сервера MG-RAST (<http://metagenomics.anl.gov/>).

Первый метагеном происходит из дренажной жидкости из металлических рудников в Калифорнии. Кислотный рудничный дренаж, который представляет экологическую опасность, формируется в результате микробной активности на сульфидных минеральных породах, подвергшихся воздействию воздуха и воды. Сообщество, которое может существовать в этих условиях низких pH, очень ограничено, и важно понять, как эти организмы создают кислый шахтный дренаж, а также, как ограниченное сообщество может быть использовано в качестве упрощенной системы, чтобы понять в общем более сложные сообщества.

Второй метагеном происходит от метагеномного обследования мирового океана доктора Крейга Вентера, новаторского исследования метагенома, который задумывался как "последовательность всего океана". Океанические среды содержат огромное разнообразие бактерий, архей и одноклеточных эукариот, а также большую морскую жизнь (рыба, кораллы и т.д.). Эта метагеномная последовательность является одной из серии 88 отдельных образцов, размещенных Вентером из его личным паруснике. Все образцы глобального обследования океана (GOS) фильтровали для обеспечения того, чтобы секвенировали только одноклеточные организмы.

Используя интерфейс MG-RAST, мы рассмотрим состав этих двух сообществ на основе последовательностей их метагенома, и сравним два образца как таксономически (кто в средах) и метаболически (какие функции происходят в этих средах).



Рис. 5. Домашняя страница сервера MG-RAST (<http://metagenomics.anl.gov/>)

1. Зайдите на MG-RAST веб-сайт: <http://metagenomics.anl.gov/> (обратите внимание, MG-RAST лучше работает с Firefox).

а) Сколько метагеномов в настоящее время размещены на сервере MG-RAST?

2. Найдите **4441138.3** с помощью функции поиска, это ID метагенома для UBA Acid Mine Drainage Biofilm. Оставьте эту вкладку открытой

3. Открыть новую вкладку и в домашней странице MG-RAST с помощью функции поиска найдите **4441147.3**. Это один из идентификаторов метагенома из "Global Ocean Sampling Expedition"..

а) Откуда была взята эта проба?



Рис. 6. Исходные данные записи метагенома. Ссылки вверху включают, где можно скачать данные о последовательностях, линк на страницу анализа MG-RAST (в красной рамке), а также различные ссылки на этот набор данных в других базах данных, в том числе NCBI

b) Для обеих записей метагеномов, прокрутите вниз до круговой диаграммы “Sequence Breakdown”.

Что содержит образец GOS с точки зрения его последовательности? Образец из кислого дренажа? Какой из образцов лучше охарактеризован на основе этих диаграмм?

Существует много информации, отображаемой на этих записях: потребуется некоторое время, чтобы посмотреть их, щелкая по разделам с правой стороны.

В разделе Taxonomic Distribution какой домен является доминирующим в каждом метагеноме? Какие филы? (подсказка - графики являются интерактивными, наведите курсор мыши на ломтики, чтобы увидеть то, что они представляют). Является ли один образец очевидно более таксономически разнообразным (много различных групп) чем другой?

Контрольный вопрос 2


Какие филы являются наиболее распространенным в пробе из Индийского океана: Cyanobacteria, Proteobacteria, Steptophyta, Nematoda, or Deinococcus?

- c) Перейдите к кривой разрежения в нижней части .Кривая разрежения отображает, был ли образец секвенирован до насыщения: крутой наклон на графике указывает на то, что образец еще не был полностью секвенирован, а наклон, который начинает выравниваться указывает на то, что большинство ДНК в пробе была секвенирована.

Сравните два графика, последовательность которого метагенома ближе к полной последовательности для этого сообщества? Ожидали ли Вы такой результат, учитывая сообщества, которые были секвенированы?

Сравнительная метагеномика

Мы использовали основные данные метагенома как грубый инструмент сравнения: сейчас давайте посмотрим на сходства и различия между этими двумя образцами более подробно.

4. Перейдите в начало Global Ocean Survey entry и нажмите на символ бар-графика. В новой вкладке откроется страница MG-RAST анализа.
5. В разделе “Data selection”кликните назеленую  кнопку следом за

“Metagenomes”. Выберите проект “Acid Mine Drainage” project из списка слева и кликните ☐. Справа появятся два новых метагенома. Выберите образец “5-way (CG) Acid Mine Drainage biofilm” и верните его в основной список, кликнув на ☐. Теперь вы должны иметь оба идентификатора метагеномов, приведенные около “metagenomes”..



Рис. 7. Страница MG-RAST анализа, где можно определить таксономические и функциональные аннотации для данных последовательностей с задаваемыми пользователем параметрами (e-value, длина выравнивания). Этот интерфейс позволяет управлять спецификой результатов. Он также позволяет проводить сравнение метагеномов непосредственно через набор инструментов MG-RAST.

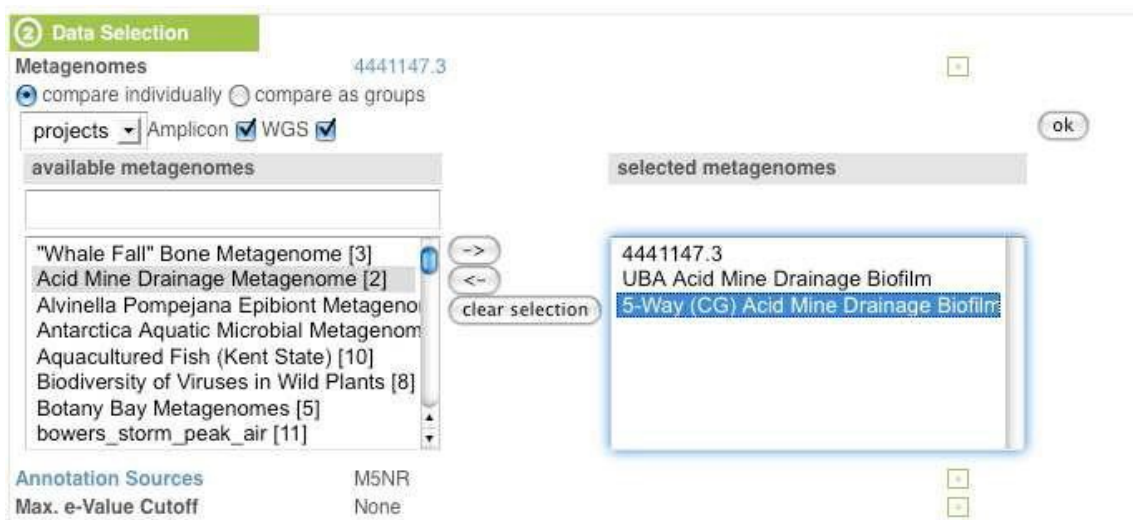


Рис. 8. Выбор метагеномов для сравнения из общедоступных данных метагеномных последовательностей MG-RAST.

6. Для сравнения двух образцов на таксономическом уровне, выберите “Best Hit Classification” в “ **Data Type**”. Затем в “ **Data Selection**” установите параметры следующим образом: измените **Max e-value cutoff** на **1e-30**, **Min. % Identity cutoff** на **90%**, и **Min. Alignment Length Cutoff** на **20**. Эти параметры определяют, какие аннотации будут приняты и будут отображаться при анализе.

- a) *Что означает изменение Min. % Identity cutoff на 80%? Почему вы хотите сделать это для вашего метагенома? Почему это может быть полезно?*
- b) В “ **Data Visualization**” выберите дерево и щелкните “generate”. Прокрутите, чтобы увидеть результат. *Какой уровень таксономии отображается?*
- c) *Какие группы окрашены?*
- d) *Что означают боксы между именами групп и диаграммой дерева?*
- e) *Что меняется, когда вы нажимаете на один из узлов Подсказка:*

если вы не видите разницы, прокрутите вправо. Что это показывает?

- f) Из дерева, сильно ли «перекрываются» сообщества двух образцов? Полностью? Назовите хотя бы одну таксономическую категорию, где представлены оба метагенома.*

Контрольный вопрос 3

Какой таксономический порядок имеют представители в обоих образцах, и в Acid Mine Drainage , и в the Indian Ocean samples?

Bacteria, Prochlorales, Burkholderiales, Spirochaetales, or Pseudomonadales?

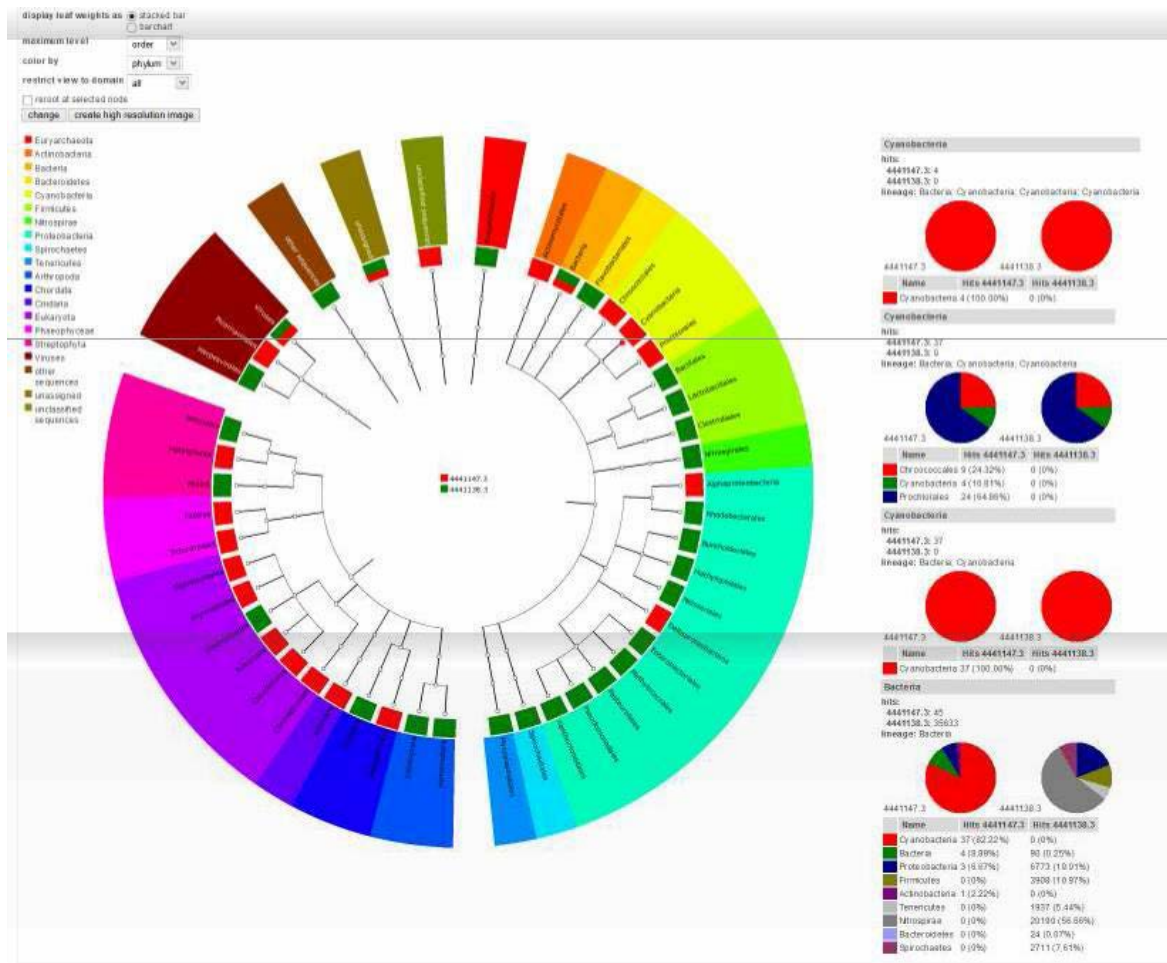


Рис. 9. Филогенетическое дерево таксономических групп, выявленных в двух метагеномах.

- Теперь давайте сравним два сообщества с функциональной точки зрения. В “**Data Type**” выберите Functional Abundance >> Hierarchical Classification. В “**Data Selection**” оставьте источник аннотаций в качестве “Subsystems”: это внутренний метод классификации в MG-RAST. Измените остальные параметры как в пункте 6. Выберите “heatmap” и щелкните.

This data was calculated for metagenomes 4441147.3 and 4441138.3. The data was compared to Subsystems using a maximum e-value of 1e-30, a minimum identity of 90 %, and a minimum alignment length of 15 measured in aa for protein and bp for RNA databases.

The heatmap was clustered using ward with bray-curtis distance metric.

group heatmap by

redraw using values, clustering and distance

The image is currently dynamic. To be able to right-click/save the image, please click the static button

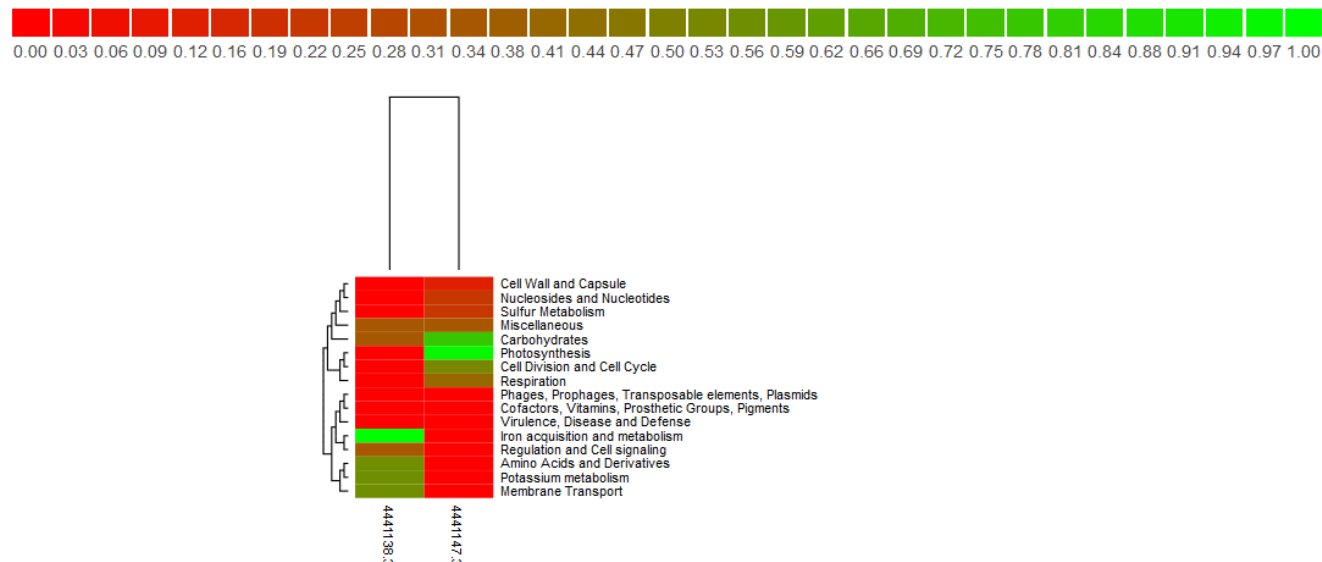


Рис. 10. Тепловая карта сравнения функциональных профилей двух метагеномов на подсистемном уровне 1.

8. Тепловая карта отображает функциональные категории и относительное обогащение белков каждого метагенома в этих категориях.

- a) Что означает красный цвет на тепловой карте ? Зеленый?
 - b) Назовите функциональную категорию, где метагеном GOS (4441147.3) является наиболее обогащенным по сравнению с сообществом Acid Mine Drainage (4441138.3). Почему это может быть? В какой функциональной категории более обогащен метагеном сообщества Acid Mine Drainage?

9. В параметрах тепловой карты изменить уровень группировки тепловой карты на уровень 2, и нажмите “draw”. Группа 2 является более

специфической схемой MG- RAST subgroups, в то время как в группе 1 очень большие, общие заголовки. (например, group 1 = “Transporters”, а group 2 within that group 1 = “methionine transporters”)

а) Какие категории самые обогащенные для каждого метабенома? Можете ли вы связать с местом, откуда был взят образец?

Полезные ссылки

MG-RAST <http://metagenomics.anl.gov/>

Приобретенные навыки и умения

- понимать некоторые современные технологии секвенирования нового поколения (NGS);
- иметь представление, как в NGS осуществляется сборка, и быть в курсе некоторых ошибок, связанных со сборкой коротких ридов и знать, как парноконцевое секвенирование может решить некоторые из этих проблем;
- знать, как исследовать RNA-Seq данные в Genome Browser и понять, что RNA-Seq трек может быть использован для установления уровня экспрессии гена, и что может происходить на уровне альтернативного сплайсинга;
- познакомиться с понятием метагеномов и понять, как они могут быть использованы для оценки видового разнообразия сообщества и метаболического потенциала организма или сообщества.

Дополнительная литература

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC. 2009. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Research* 20: 45–58.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. 2008. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386.

Nakano M, Nobuta K, Vemaraju K, *et al.* 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Research* 34:D731-5.